Statistical Data Mining | Classification and GLM

Pablo X Zumba

1) Create three separate models to understand the predictors of churn:
    (i) subscribers of telephone services
    (ii) subscribers of internet services
    (iii) people who subscribe to both services

2) What predictors do you think contribute to the churn of (i) only telephone customers, (ii) only Internet service customers, and (iii) customers who subscribe to both phone and Internet services? Explain the rationale for your answer.

3) Create training and test data sets with a 75:25 split using a random seed of the last 4 digits of your U-number to set the random split. Use the training data to train three logit models with the variables you identified in Question 2. Combine the outputs of the three models using stargazer.

4) What are the top three predictors of churn of (i) only telephone customers, (ii) only Internet service customers, and (iii) customers who subscribe to both phone and Internet services? Explain using marginal effects how much each predictor contributes to churn occurrence.

5) Fit your models using test data, and compute recall, precision, F1-score, and AUC values for each of your three models. Create a table with these values.

Analyze the data carefully (data definitions are provided in the second worksheet of the Excel file). Submit your results in the form of a nicely formatted Word (or PDF) file and your R code as two separate files.

Table of relevant predictors hypothesized direction of effect (+/-), and the rationale for each hypothesized effect.

Predictor	Telephone	Internet	Both Services	Rationale for effect
Churn - Predictor				
SeniorCitizen	Yes	Yes	Yes	I think they are more susceptible to feeling cheated.
tenure	Yes	Yes	Yes	Yes, because it is linked to customer loyalty.
PhoneService	Yes	Yes	Yes	Since some services are linked to some functionalities, not having a good telephone service increases churn on the internet and vice versa.
MultipleLines	Yes	No	No	Since there may be discounts, the more lines, the less likely you are to churn.
InternetService	No	Yes	No	Since some services are linked to some functionalities, not having a good telephone service increases churn on the internet and vice versa.
OnlineSecurity	No	Yes	No	Having a bad online security service increases the chance to churn on the internet.
OnlineBackup	No	Yes	No	Having a bad online backup service increases the chance to churn on the internet.

TechSupport	Yes	Yes	Yes	Having a bad Tec support service could increase the chance to churn on all services due to customer experience.
StreamingTV	No	Yes	No	Churn could result from a poor streaming service.
StreamingMovies	No	Yes	No	Churn could result from a poor streaming service.
Contract	Yes	Yes	Yes	Customer churn could result from an inflexible contract policy.
MonthlyCharges	Yes	Yes	Yes	Churn can occur if charges do not meet customers' budgets.
TotalCharges	Yes	Yes	Yes	Churn can occur if charges do not meet customers' budgets.
Excluded: **customerID** (Not useful due to there is no information related to customers behavior), **gender** (irrelevant), **partner(**irrelevant**),** **dependens** (reverse effect due to it is more likely to subscribe to any services if have dependents), **deviceprotection**(irrelevant), **PaperlessBilling**(Not relevant for the prediction), **PaymentMethod** (Not relevant for the prediction).				

Models for each service/scenario:

Telephone:
- **logit_telephone = glm(churn ~ tenure+contract+totalcharges+seniorcitizen, family=binomial (link="logit"), data=churn_telephon e)**
- **probit_telephone <- glm(churn ~ tenure+contract+totalcharges+seniorcitizen, family=binomial (link="probit"), data=churn_telephone)**

Internet:
- **logit_internet = glm(churn ~ tenure+contract+dependents+seniorcitizen, family=binomial (link="logit"), data=churn_internet)**
- **probit_internet = glm(churn ~ tenure+contract+dependents+seniorcitizen, family=binomial (link="probit"), data=churn_internet)**

Both Services:
- **logit_both = glm(churn ~ tenure+contract+onlinesecurity+totalcharges+onlinebackup, family=binomial (link="logit"), data=churn_both)**
- **probit_both = glm(churn ~ tenure+contract+onlinesecurity+totalcharges+onlinebackup, family=binomial (link="probit"), data=churn_both)**

Churn for Telephone services

	Dependent variable:		
	churn		
	OLS	*logistic*	*probit*
	(1)	(2)	(3)
tenure	-0.010	-0.114***	-0.055***
		(0.007)	(0.003)
contractOne year		-1.114***	-0.639***
		(0.107)	(0.059)
contractTwo year		-2.285***	-1.144***
		(0.180)	(0.085)
totalcharges	0.0001	0.001***	0.0005***
		(0.0001)	(0.00003)
seniorcitizen	0.155	0.546***	0.334***
		(0.081)	(0.048)
Constant	1.459	0.164***	0.050
		(0.051)	(0.031)
Observations	6,352	6,352	6,352
Log Likelihood		-2,827.021	-2,853.012
Akaike Inf. Crit.		5,666.043	5,718.025
Note:	*p<0.1; **p<0.05; ***p<0.01		

Churn for internet services

	Dependent variable:		
	churn		
	OLS	*logistic*	*probit*
	(1)	(2)	(3)
tenure	-0.004	-0.029***	-0.016***
		(0.006)	(0.003)
contractOne year	-0.191	-1.135***	-0.649***
		(0.326)	(0.175)
contractTwo year	-0.192	-2.410***	-1.173***
		(0.621)	(0.266)
dependentsYes	-0.095	-0.791***	-0.475***
		(0.267)	(0.150)
seniorcitizen	0.131	0.706***	0.407***
		(0.261)	(0.155)
Constant	1.476	0.128	0.051
		(0.159)	(0.096)
Observations	680	680	680
Log Likelihood		-293.631	-294.845
Akaike Inf. Crit.		599.261	601.691
Note:	*p<0.1; **p<0.05; ***p<0.01		

Churn for both services

	Dependent variable:		
	churn		
	OLS	*logistic*	*probit*
	(1)	(2)	(3)
tenure	-0.013	-0.119***	-0.062***
		(0.008)	(0.004)
contractOne year	-0.157	-0.903***	-0.512***
		(0.116)	(0.065)
contractTwo year	-0.179	-2.004***	-1.020***
		(0.197)	(0.099)
onlinesecurityYes	-0.106	-0.674***	-0.386***
		(0.086)	(0.050)
totalcharges	0.0001	0.001***	0.001***
		(0.0001)	(0.00004)
onlinebackupYes	-0.044	-0.281***	-0.165***
		(0.080)	(0.047)
Constant	1.605	0.685***	0.388***
		(0.059)	(0.036)
Observations	4,832	4,832	4,832
Log Likelihood		-2,411.304	-2,423.487
Akaike Inf. Crit.		4,836.608	4,860.974
Note:	*p<0.1; **p<0.05; ***p<0.01		

Interpretation for telephone services models:

- For a one-unit increase in tenure, the Churn increases by 0.011 in the logit model and 0.055 in the probit model for telephone services.
- For a one-unit increase in total charges, the Churn increases by 0.001 in the logit model and 0.0005 in the probit model.
- It is more likely that all services will end up in a churn if a contract lasts for a long time, i.e., there is less probability to churn if the contract is month to month and the highest probability if the contract is two years.
- Senior citizens are more likely to churn from internet services than from telephone services.

Point out the top predictors for each dependent variable. Explain using marginal effects how much each predictor contributes to churn occurrence.

- **Telephone service:** tenure, contract, total charges, and senior citizen.
- **Internet service:** tenure, contract, dependents, and senior citizen.
- **Both services:** tenure, contract, online security, and total charges.

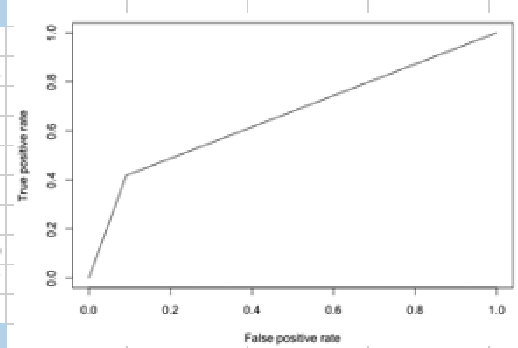
Out-of-sample exercise: carefully explain how you tuned the classifiers. Compare the prediction metrics across models.

The Recall, Specificity, Precision, Accuracy, F1 Score, Misclassification/Error Rate, Prevalence, and AUC were calculated using the formulas given a Confusion matrix as input values.

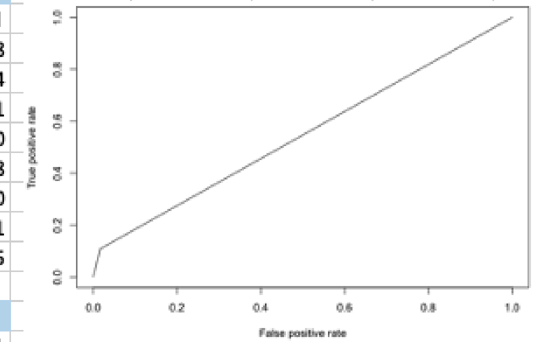
The classifiers were tuned as follows:

1) Subset/filter the data for each scenario, telephone, internet, and both services.
2) Run a first model which includes all the variables that make sense and that the model allows, so in the first scenario for telephone services, most of the independent variables were used except those related to internet service.
3) Import the library "caret" to run the function "varImp" on the previous model's output. A vector will be returned with the most important variables labeled by their highest number.
4) Choose the first 4-5 variables with the highest number and run the model again using those variables.
5) For the other two scenarios, follow the same procedure.

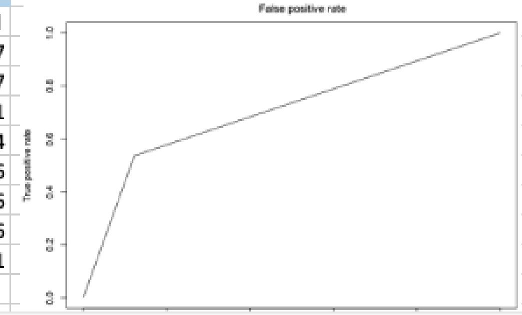
Confusion Matrix of test Churn Telephone data.								
		TP	FN				Formula	CalculatedVal
		FP	TN			Recall/Sensitivity	=TP/Predicted Yes	0.418
						Specificity	=TN/Predicted NO	0.909
						Precision	=TP/Actual Yes	0.631
Total =	2268	**Predicted**				Accuracy	=(TP+TN)/Total	0.775
	Actual	258	359	*Predicted YES =*	617	F1 Score	=2/((1/Recall)+(1/Precision))	0.503
		151	1500	*Predicted NO =*	1651	Misclassification/Error Rate	=(FP+FN)/Total	0.225
		Actual YES=	*Actual NO=*			Prevalence	=Actual YES/Total	0.180
		409	1859			AUC	Calculated in R	0.67



Confusion Matrix of test Churn Internet data.								
		TP	FN				Formula	CalculatedVal
		FP	TN			Recall/Sensitivity	=TP/Predicted Yes	0.108
						Specificity	=TN/Predicted NO	0.984
						Precision	=TP/Actual Yes	0.711
Total =	6522	**Predicted**				Accuracy	=(TP+TN)/Total	0.750
	Actual	189	1555	*Predicted YES =*	1744	F1 Score	=2/((1/Recall)+(1/Precision))	0.188
		77	4701	*Predicted NO =*	4778	Misclassification/Error Rate	=(FP+FN)/Total	0.250
		Actual YES=	*Actual NO=*			Prevalence	=Actual YES/Total	0.041
		266	6256			AUC	Calculated in R	0.55



Confusion Matrix of test Churn Both services data.								
		TP	FN				Formula	CalculatedVal
		FP	TN			Recall/Sensitivity	=TP/Predicted Yes	0.537
						Specificity	=TN/Predicted NO	0.877
						Precision	=TP/Actual Yes	0.621
Total =	3408	**Predicted**				Accuracy	=(TP+TN)/Total	0.784
	Actual	499	431	*Predicted YES =*	930	F1 Score	=2/((1/Recall)+(1/Precision))	0.576
		304	2174	*Predicted NO =*	2478	Misclassification/Error Rate	=(FP+FN)/Total	0.216
		Actual YES=	*Actual NO=*			Prevalence	=Actual YES/Total	0.236
		803	2605			AUC	Calculated in R	0.71



Explanation of performance Metrics:

- AUC = 0.71 indicates that the best model fit occurred when predicting both services.
- While the accuracy was high when predicting internet data (= 0.75), the AUC was just 0.55, as there were substantially fewer observations when compared with the telephone subset data and both subsets.
- While both services have a precision of only 0.621 when predicting data, other performance metrics such as accuracy and specificity make the AUC higher.
- In “both services data”, the best-predicting model has the lowest misclassification rate, so it is the best-predicting model.